Text Line Segmentation using Vector Quantization Incorporated Cluster Density Analysis

Asha B M[1], Purushotham U[2]
[1] Department of ECE, PESIT, Bangalore, India
[2] Assistant Professor, Department of ECE, PESIT, Bangalore, India

***Abstract*: This paper proposes a novel technique of text line segmentation. The method is based on the concept of application of vector quantization strategy and cluster density analysis. This approach is able to detect text lines in handwritten image document which may contain lines oriented in different direction and varying space between main lines.**

Keywords:* *Vector quantization, cluster density, line segmentation

I. INTRODUCTION

“Vector quantization Strategy Incorporated cluster density Analysis: An efficient approach for text line segmentation [1] [2] [4] considers the application of the statistical methods to determine the probability of identifying a line. The methods primarily rely on the assumption that the density of the cluster within one region has equivalent values.

The methodology is constituted of two phases. In the first phase the image is decomposed into its components representing an encapsulation of the text content. The next phase involves grouping mechanism that is a classical quantization technique based on the modeling of probability density functions by the distribution of prototype vectors.

The vector quantization technique is used to capture the similarity or the distance between the applicable samples. This technique aids in knowing distance between the samples for grouping them into cluster. This grouping them into cluster is done using cluster density analysis.

The cluster density [3] analysis works by organizing the components to form a group having approximately the same number of points closest to them across the components. From the group, the line segment is constructed by linking the components one by one to form line cluster. At each step we accumulate one additional component to form a cluster. Thus, first cluster is one with two components that have the shortest distance. The next component to be added shall be the one with the next least distance, the process is continued till a line is formed. The entire step is iterated till all the components have been segmented into meaningful lines.

In this paper section II discuses about proposed block diagram, where individual blocks are explained in detail. In section III results of handwritten bmp file is shown. Finally in section IV we conclude the paper.

II. PROPOSED BLOCK DIAGRAM

The block diagram is shown in figure 1. The core of the logic is the componentization and the line segmentation block. The algorithm exhibits high accuracy of text line segmentation as well as delivers an improved performance.

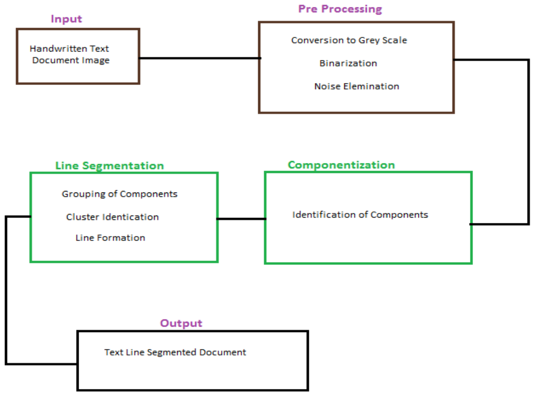


Figure 1: Block diagram

Input: Input defines the mode in which the documents/ images containing the text are acquired from input devices. Typically these images are in the format of some picture files as BMP, TIFF, PNG or JPG.

***Pre-Processing*:** Pre-processing constitutes the line of action needed before segmentation. These include binarization, noise filtering and optional thinning operation. Binarization process converts the input image which is in the RGB format into a grey scale image and finally into its binary equivalent format. In this process each pixel value is compared with its threshold, it is retained the value as 1 or else the value is made 0. The noise filtering process removes the noise from the image that may have appeared during the scanning process or due to soiling of the document. Optionally process of thinning is also applied to reduce the width of the line in the images.

***Componentization*:** The componentization [3] [8] process involves a two pass algorithm to find the connected components and label them. We also find the centroid of all the components so formed. The text line segmentation logic selects a set of components to form a group based on the vectored density of each component. The nearest neighboring logic is then applied within the cluster to link the components. The sorting formula is then applied to form a sequential text line component.

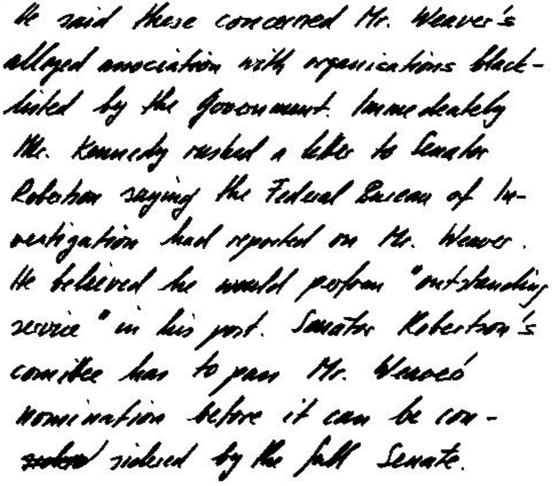
Segmentation: Segmentation [5] [6] is a process that decomposes an image of sequence of characters into sub images of individual symbols. Skew detection and correction or block identification forms a pre-step to line segmentation. The skew estimation defines the angle that the line of text in the digital image is inclined in comparison to the horizontal. Line segmentation is a technique of bifurcating the image context into individual lines. Word segmentation separates the words from the lines obtained and the method generally used is vertical projection profile technique. Character segmentation refers to splitting of words into characters.

Recognition: The recognition [7] part involves feature extraction and classification. Every character shall have a specific feature and in this stage OCR systems analysis and selects a feature that can be used to uniquely identify the character segment. Classification stage is the main decision making stage of the system and uses the features extracted in the previous stage to identify the text segment as per the predefined rules.

III. RESULTS

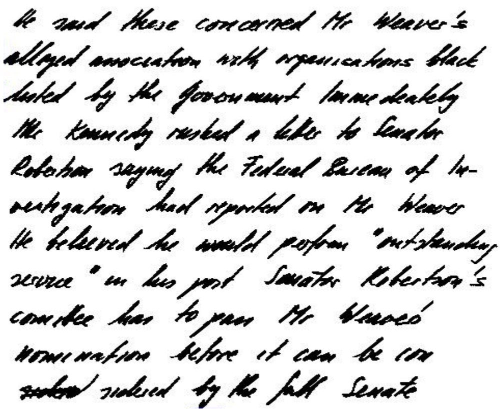
The gray scale input image in bmp format for text line segmentation is shown under the heading gray scale image. This image after elimination of noise is shown under the heading image after noise elimination. This noise eliminated image is clustered into components by componentization technique. Thus obtained image is used for identifying the lines in the image. The output after identifying the line is shown under the heading final output.

Gray scale image:



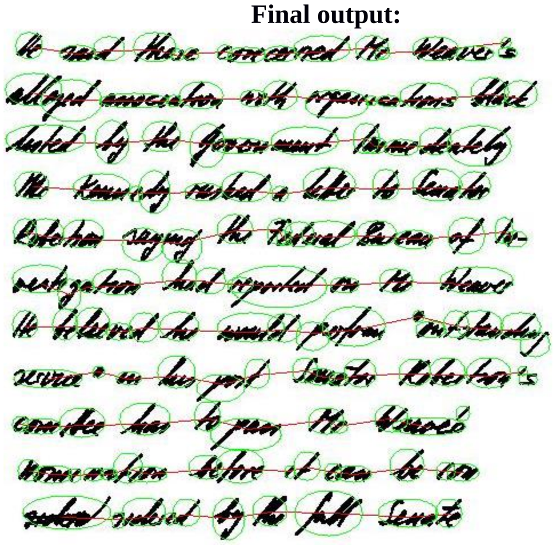
He said these concerned Mr. Weaver's
alleged association with organizations black-
listed by the Government. Immedeately
Mr. Kennedy rushed a letter to Senator
Robertson saying the Federal Bureau of In-
vestigation had reported on Mr. Weaver.
He believed he would perform "outstanding
service" in his post. Senator Robertson's
committee has to pass Mr. Weaveŕs
nomination before it can be con-
sidered by the full Senate.

Image after noise elimination:



He said these concerned Mr Weaver's
alleged association with organizations black
listed by the Government Immedeately
Mr Kennedy rushed a letter to Senator
Robertson saying the Federal Bureau of In-
vestigation had reported on Mr Weaver
He believed he would perform "outstanding
service" in his post Senator Robertson's
committee has to pass Mr Weaveŕs
nomination before it can be con
sidered by the full Senate

Final output:



He said these concerned Mr Weaver's
alleged association with organizations black
listed by the Government Immedeately
Mr Kennedy rushed a letter to Senator
Robertson saying the Federal Bureau of In-
vestigation had reported on Mr Weaver
He believed he would perform "outstanding
service" in his post Senator Robertson's
committee has to pass Mr Weaveŕs
nomination before it can be con
sidered by the full Senate

IV. CONCLUSION

The novel method proposed in this paper for text line segmentation is vector quantization and cluster density analysis. This method is able to recognize the handwritten text lines correctly. Even though the written text line letters are overlapping, these methods will correctly identify the letters belonging to the corresponding lines. This method is able to detect to correct text lines up to 98%. Inspite of the drawbacks that is present, this method is able detect correct text lines.

REFERENCES

[1] Z. Razak, K. Zulkiflee, et al., Off-line handwriting text line segmentation: a review, International Journal of Computer Science and Network Security 8 (7) (2008) 12–20.

[2] Off-line Handwriting Text Line Segmentation: A Review, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008, University of Malaya, Kuala Lumpur, Malaysia

[3] Handwritten Chinese text line segmentation by clustering with distance metric learning, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences.

[4] Handwritten Text Line Segmentation by Shredding Text into its Lines, 2009 10th International Conference on Document Analysis and Recognition.

[5] Text line and word segmentation of handwritten documents, Received 8 August 2008, Revised 12 November 2008, Accepted 21 December 2008.

[6] Segmentation of off-line cursive handwriting using linear programming by BERRIN YANIKOGLU, PETER A. SANDON, received 16 November 1995, revised 22 May 1998, accepted 22 May 1998.

[7] G. Louloudis, B. Gatos, C. Halatsis, Text line detection in unconstrained handwritten documents using a block-based Hough transform approach, in: Proceedings of International Conference on Document Analysis and Recognition, 2007, pp. 599–603.

[8] C.-H. Chou, S.-Y. Chu, F. Chang, Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition 40 (2) (2007) 443–455.